David P. Baker, Senior Research Scientist
Casey Mulqueen, Research Scientist
American Institutes for Research
3333 K Street, NW
Washington, DC 20007

R. Key Dismukes, Chief Scientist
Aerospace Human Factors Research Division
NASA Ames Research Center
Moffett Field, CA 94035

TRAINING PILOT INSTRUCTORS TO ASSESS CRM: THE UTILITY OF FRAME-OF-REFERENCE (FOR) TRAINING

ABSTRACT

The extent to which pilot instructors are trained to assess crew resource management (CRM) skills accurately during a simulator scenario is critical. Pilot instructors must make accurate performance ratings to ensure that proper feedback is provided to the flight crew and appropriate decisions are made regarding certification to fly the line. This paper reviews several approaches to rater training and identifies what we believe would be the most effective approach for training
pilot instructors to assess CRM: Frame-Of-Reference (FOR) training. The goal of FOR training is to train pilot instructors to common standards, which are developed by expert instructors. Research suggests that if pilot instructors were trained to evaluate performance using the same standards as "experts" they should produce more accurate ratings. Based on the results of this research, specific guidelines are presented for developing FOR training and the benefits and limitations of this training are discussed. Finally, we conclude with a series of unanswered questions regarding pilot instructor rater training that require investigation within the airline industry.

INTRODUCTION

Crew resource management (CRM) training has long been a concern in commercial aviation[1]. This training was developed to address the problem of human error on the flight deck. Since its inception, CRM training has evolved significantly and is now an integral component of pilot training conducted under the Advanced Qualification Program (AQP) as described in SFAR58[2] and FAA Advisory Circular AC 120-54[3]. AQP provides a voluntary alternative to traditional pilot training under CFR 14, Parts 121 and 135. AQP integrates CRM concepts throughout the pilot training curriculum and requires pilots to demonstrate proficiency in scenarios that test both technical and CRM skills together prior to certification. Currently, most major air carriers as well as a growing number of regional airlines participate in AQP.

The introduction of AQP has led to a significant amount of research on the process by which pilots are evaluated on their CRM skills. Under AQP, pilots are trained and assessed during Line Operational Simulation (LOS) scenarios (i.e., Line Oriented Flight Training [LOFT], Line Operational Evaluation [LOE], and

Special Purpose Operational Training [SPOT]) during initial or recurrent training. LOFT and SPOT are used for CRM training while LOE involves actual evaluation of the flight deck crew's CRM skills. All LOS scenarios involve a complete cockpit crew (i.e., Captain, First Officer, and Flight Engineer [depending on aircraft type]) flying a scenario in a full motion simulator. These scenarios usually begin at the departure gate and include specific scenario events that are introduced as the flight progresses to the destination airport. Each scenario event set is designed to elicit technical and CRM behaviors by the crew[4]. A pilot instructor, seated in the back of the simulator, observes the crew's response to each event set and rates the performance of the crew and each crewmember regarding their technical and CRM skills.

A critical element in LOS is the pilot instructor (the term "pilot instructor" is employed throughout this paper, but it should be noted that this term encompasses any qualified individual directly involved in training and evaluating an aircrew's CRM skills during LOS [e.g., pilot instructors, check airmen, Standards Captains, etc.]). As noted in the preceding paragraph, these individuals observe how each aircrew performs on the LOS scenario event sets and assigns technical and CRM performance ratings[5]. In LOE, the resulting ratings are used to determine whether or not each pilot in the crew should be certified to fly the line or requires additional training prior to certification. Therefore, the extent to which pilot instructors make accurate judgments about crew and crewmember performance is critical to the effectiveness of AQP training and airline operations.

A reliable and accurate assessment of a crew's CRM skills can not be made during a LOS scenario if pilot instructors do not agree on the CRM behaviors observed and the level of performance demonstrated for each skill. When pilot instructors do not agree, performance ratings are a function of the particular instructor conducting the LOS as opposed to performance of the crew. To safeguard against this problem, Longridge and others have suggested that pilot instructors should receive formal rater training[5,6]. Under AQP, pilot instructor training is required and the proficiency and standardization of instructors and evaluators must be verified on a recurrent basis[2,3].

Given the critical role pilot instructors play in LOS, the purpose of this paper is to identify an effective strategy for training pilot instructors to be accurate when assessing CRM (i.e., pilot instructor rater training). We do this by first reviewing four strategies from the field of performance appraisal that have been used traditionally to train supervisors to be reliable and accurate when making performance assessments. These strategies are directly generalizable to pilot instructor rater training. In addition, we review research on each strategy's effectiveness to determine the best approach for training pilot instructors. Next, we will discuss how several air carriers have trained pilot instructors under AQP and draw comparisons between this approach and what the research on performance appraisal suggests are the "best practices" for training raters. Finally, we present a set of specific guidelines for developing pilot instructor rater training. These guidelines describe specific content that should be included in any pilot instructor rater training program.

STRATEGIES FOR TRAINING PILOT

INSTRUCTORS

The vast majority of the research that is relevant for training pilot instructors to make accurate and reliable ratings when assessing CRM has been conducted in the domain of performance appraisal[7]. In performance appraisal, a supervisor observes and evaluates subordinate performance on a number of job-related dimensions. This process is similar to a LOS scenario except that LOS involves the assessment of an aircrew rather than an individual, and the assessment occurs during a defined scenario rather than over an extended performance period. Therefore, we believe that the lessons learned from performance appraisal are directly generalizable to LOS and could be leveraged to construct effective pilot instructor rater training.

Historically, four strategies have been advocated for training raters to be accurate when making performance judgments. These are: Rater Error Training (RET), Performance Dimension Training (PDT), Behavioral Observation Training (BOT), and Frame-Of-Reference (FOR) training. With the exception of BOT, each of these approaches has been widely studied. In this section, we briefly describe each of these training strategies. In the section that follows, we present empirical evidence pertaining to the effectiveness of each.

Rater Error Training (RET).

The purpose of RET is to familiarize raters with common rating errors in hopes that such knowledge will reduce these errors and produce more accurate ratings. RET is accomplished by providing raters with a detailed lecture about common rating errors that can occur during performance evaluation. These include halo error (i.e., the tendency of a global impression of a ratee [i.e., aircrew, subordinate, etc.] to dictate ratings on all performance dimensions), leniency error (i.e., the tendency to give ratees high performance ratings), severity error (i.e., the tendency to give ratees low performance ratings) and central tendency error (i.e., the tendency only to give ratees performance ratings near the middle of the performance scale). In all cases, the rater making the performance ratings fails to distinguish among different performance levels and typically clusters ratings within one part of the rating scale. Therefore, the desired outcome of RET is performance ratings that are more normally distributed.

Performance Dimension Training (PDT).

The purpose of PDT is to familiarize raters with the rating scales that will be used to evaluate different dimensions of performance. This training is usually accomplished by having raters review and discuss the rating scales or involving raters in the actual development of the scales. PDT is based on research that suggests that people tend to form evaluative judgments at the time that behavior is observed rather than at a later time when making performance ratings[8]. Therefore, PDT trains raters to recognize and use the appropriate performance dimensions and to rely upon these dimensions when making observations. As a result, performance ratings should be based on behavior that was observed and organized by job-related dimensions producing more accurate ratings[7,9].

Behavioral Observation Training (BOT).

The purpose of BOT is to increase the observational skills of raters. Unlike RET and PDT, BOT focuses on the observation of behavior rather than the evaluation or rating of behavior. BOT is based on the premise that there is a significant difference between the processes involved in observation, and those involved in evaluation[10]. According to this view, observation processes encompass the detection, perception, and recall of behavioral events, while evaluation processes include categorizing, integrating, and evaluating information. In BOT, faulty behavioral observation is viewed as the primary reason for rating inaccuracies. Typically, BOT encompasses strategies that focus on the observation or recording of behavior (e.g., note-taking, diary-keeping, etc.). Discussion and/or practice exercises that focus on recognizing and avoiding systematic errors of observation, contamination from prior information, and overreliance on a single source of information may also be included[11].

Frame-Of-Reference (FOR) Training.

Finally, the purpose of FOR training, as the name implies, is to train raters to a common frame-of-reference[12]. Here, rater trainees are presented with information about the rating task and the relevant performance dimensions to be assessed. Raters are given samples of varying levels of performance on behaviors that represent each dimension, along with practice and feedback in the use of these performance standards[13]. The defining characteristic of FOR training is the nature of practice and feedback provided to rater trainees. Here, practice usually involves rating a series of training videotapes that present varying levels ratee performance, and feedback usually compares a rater trainee's practice ratings to a set of previously defined "true scores." True scores are assigned to each videotape by experts who review the tape, independently rate the performance, and discuss their ratings to reach consensus. The resulting ratings are believed to reflect the actual performance level displayed on the videotape[14,15].

Effectiveness of Rater Training

The most comprehensive summary of the research on rater training is a meta-analytic review of twenty-nine studies from the field of performance appraisal conducted by Woehr and Hoffcutt[7]. Meta-analysis reports results using an effect size statistic, *d*, which, in this case, represents the effectiveness of a rater training method. Therefore, a positive *d* value indicates that the training was effective, while a negative *d* value indicates the opposite. A *d* of .2 indicates a small effect, a *d* of .5 represents a medium effect, and a *d* of .8 represents a large effect[16].

For the purpose of the meta-analysis, each rater training strategy was analyzed against measures of observation and rating accuracy to determine effectiveness. Observation accuracy is related to the extent to which raters can correctly identify and record ratee behaviors, while rating accuracy is related to the extent to which raters can assign the appropriate performance ratings (i.e., on a defined rating scale) to the behaviors that were observed. In both cases, observations and/or ratings are compared to "true scores" derived by task experts to determine accuracy.

Turning to the meta-analysis results, FOR training was found to be the *most effective* strategy for increasing rating accuracy, with

a mean effect size of d = .83. FOR training was also found to have a positive effect (d = .37) on observation accuracy. BOT was found to be almost as effective as FOR in training rating accuracy (d = .77) and somewhat better in improving observation accuracy (d = .49). However the results for BOT should be viewed with caution since results from only four studies were available for the meta-analysis. Regarding PDT, the results showed that this strategy had a weak positive effect on rating accuracy (d = .13). However, no data were available on the effects of PDT on observation accuracy. Finally, RET was found to have a slight positive effect on rating accuracy (d = .26) and a slight *negative* effect on observation accuracy (d = -.17)[7]. A similar review of the effectiveness of different rater training strategies (RET, PDT, and FOR) by Smith produced comparable results[17].

In addition to examining the effectiveness of RET, PDT, BOT, and FOR, Smith reviewed the research on the effectiveness of different rater training methods (lecture, group discussion, and practice and feedback). This review found that the inclusion of practice and feedback is critical for improving the accuracy of raters. Increases in accuracy were reported in five out of the six studies reviewed by Smith that included practice and feedback. Only one study was reported that utilized discussion alone and it failed to result in any increases in rating accuracy. Regarding the lecture method, this approach to rater training was generally associated with RET and for the majority of studies found to be ineffective. Five out of the eight studies reviewed found that lecture failed to improve rater accuracy. Of the studies that reported an increase in rating accuracy, lectures were either combined with practice and feedback, or discussion and practice and feedback[17].

In summary, several conclusions can be drawn from the literature cited above that have direct relevance for training pilot instructors to assess CRM skills. First, of the various training strategies reviewed, FOR training produced the greatest increases in
rating accuracy. Although FOR has not yet been tested in the specific area of training pilot instructors to assess CRM performance in LOS scenarios, we believe it would be effective and we are conducting research to evaluate this approach. Essentially, pilot instructors trained to evaluate CRM skills using expert instructor standards (i.e., what have been referred to as "gold standards" in the airline industry) should produce ratings more like these experts. Second, although BOT represents a relatively new and unstudied methodology, it appears to be effective for increasing both observational and rating accuracy. Because the evaluation of CRM skills requires pilot instructors to make behavioral observations of flight deck crews on each event set that comprises a LOS scenario, we recommend that training in observational skills be included in any pilot instructor rater training program. Finally, the literature suggests that the combination of group discussion with significant opportunities for practice and feedback

Table 1. Pilot instructor rater training best practices.

Best Practices
1. Presentation and discussion of the of the CRM skills to be rated. 2. Discussion of the standards associated with each CRM skill to be assessed. 3. Training on behavioral observation. 4. At least three opportunities to practice the

rating task on videotapes of aircrew flying specific LOS scenario event sets.
5. Feedback that compares pilot instructor practice ratings to gold standards.

is the most effective training approach. Although the amount of practice and feedback required has yet to be determined, we recommend that pilot instructor rater training programs include at least three opportunities (i.e., that span a poor to excellent performance) to practice and receive feedback on the CRM evaluation task. Table 1 summarizes these conclusions in a series of "best practices" for training pilot instructors to assess CRM during LOS. These practices are primarily based on FOR training, supplemented by BOT, and extended to pilot instructor rater training.

RATER TRAINING IN THE AIRLINE INDUSTRY

In the airline industry, Interrater Reliability (IRR) training has been used at several US air carriers to train pilot instructors to assess CRM skills during LOS scenarios[6]. IRR training usually consists of a one-day workshop in which pilot instructors receive information and discuss aspects of the LOS scenario rating process and practice rating the videotaped performance of several crews. Regarding the practice component of IRR training, a videotape of a crew flying a specific LOS scenario, or one of the scenario's component event sets, would be shown to a class of instructors. These individuals then independently rate the crew's CRM performance. During a class break, ratings are analyzed to determine the current level of agreement that exists across pilot instructors and areas where significant rating discrepancies exist. Upon reconvening the class, the results of these analyses are fed back to the workshop participants and rating discrepancies are discussed to reach consensus. Videotape of a different crew flying the same LOS scenario is then rated to determine the level of agreement achieved within the class[6].

To date only a handful of studies have been conducted on the effectiveness of IRR training and the results that have been reported are mixed. In some cases IRR training has been found to increase observational and rating accuracy while in other cases no noticeable effects were found[18]. In fact, on a number of occasions raters showed high levels of interrater reliability initially and therefore no training effect was observed. In light of these mixed results, we suggest that more research needs to be conducted to determine IRR training's true effectiveness.

In absence of empirical evidence, the effectiveness of IRR training can be examined by comparing the training strategies employed in IRR to the pilot instructor rater training "best practices" described in Table 1. Referring to Table 1, IRR training includes most of the effective strategies listed.
There is discussion of the technical and CRM skills to be rated, there is discussion of the technical and CRM standards associated with each LOS scenario, and there are opportunities to practice and receive feedback on the rating task. However, no training in observational skills is presently included and feedback is based on the extent to which pilot instructors agree with each other (i.e., a group standard) rather than on the extent to which instructor ratings agree with an expert gold standard. Given these differences, it seems possible that IRR training could lead to pilot instructors

calibrated to the group standard established in their IRR training classes, but separate IRR classes might not agree with each other.

PILOT INSTRUCTOR RATER TRAINING GUIDELINES

The empirical research presented in this paper from the field of performance appraisal demonstrates the potential effectiveness of FOR training for training pilot instructors to accurately assess CRM skills during LOS scenarios. As was described earlier, FOR training is distinguished by the following characteristics:

- Discussion of the standards associated with each CRM skill to be assessed;
- Opportunities to practice and receive feedback on the CRM rating task; and
- Feedback that compares pilot instructor practice ratings to expert instructor gold standards.

In this section, we rely upon these characteristics and the best practices that appear in Table 1 to present a series of guidelines for developing pilot instructor rater training that is based on the FOR framework. Currently, these guideline are being utilized in the development of pilot instructor training at a major U.S. air carrier. Once developed, the pilot instructor rater training program will be implemented and tested at the carrier.

Guideline 1. Pilot instructor rater training should include a detailed review of the LOS scenarios.

A detailed review of the LOS scenarios to be evaluated should be included in pilot instructor rater training. Although this is not a defining feature of FOR training, this practice is important. In addition to the LOS scenario, the review should cover the event sets that comprise the LOS and the CRM skills to be evaluated. In cases where pilot instructors are being trained for the first time, this review should also include a detailed explanation of the grade sheets. Review of the various ratings to be made and any grading rules that apply (e.g., cases where certain behavioral observations lead to specific performance ratings on the CRM skills assessed). In cases where pilot instructors are receiving recurrent rater training, changes to the grade sheet or the grading process should be noted and discussed.

Guideline 2. Pilot instructor rater training should include a review and discussion of the performance standards associated with each CRM skill to be rated.

In addition to a review of the LOS scenario and the CRM skills to be assessed, pilot instructor rater training should include a review of the performance standards for each CRM skill to be rated. This review is the first step in developing consistent standards across pilot instructors for evaluating CRM skills during LOS. Information regarding the CRM requirements for successful performance of each LOS event set is often found in or can be developed from the scripts that describe the LOS scenario. Information from the LOS scripts could be leveraged to develop specific examples of different performance levels on the grade sheet. For most air carriers, examples of “Excellent”, “Standard”, “Debriefed”, and “Repeat” levels of performance would be developed. Pilot instructors could then use these examples as referents during the LOS rating process, which should enhance CRM rating accuracy and reliability.

Guideline 3. Pilot instructor rater training should train instructors to be good observers.

In order to rate crews accurately and meaningfully, instructors must be able to observe the relevant behaviors of the crew and interpret those behaviors appropriately. Accurate ratings can not occur without accurate observations. Therefore we believe that observation training should be included as part of pilot instructor FOR rater training. Unfortunately, we are not aware of any research on how to teach instructors to notice and interpret relevant crew behaviors. Anecdotal reports from airline training departments suggest that instructors vary greatly in what aspects of a given crews' performance they notice and also vary greatly in how they interpret what they notice.

Observation training should include both a discussion and a practice and feedback component. First, discussion should focus on the nature of a good observation (i.e., specific, behavioral, verifiable, etc.) and how to accurately observe an aircrew's performance during LOS. Discussion may be particularly beneficial in recurrent pilot instructor rater training, because pilot instructors could share their experiences regarding observation strategies that they have found to be effective and ineffective. Second, observation training should include opportunities for practice and feedback. The research on rater training suggests that practice and feedback is critical for training transfer[17]. Therefore, pilot instructors should be shown a series of videotapes for the purpose of practicing their observational skills. The videotapes should be annotated with detailed observations from experts about the specific behaviors exhibited by the crews and how those behaviors are best interpreted. This annotation provides detailed feedback to the instructors so they can compare what they observed or failed to observe and how they interpreted their observations to observations and interpretations of experts.

Guideline 4. Pilot instructor rater training should include opportunities to practice and receive feedback on the rating task.

Relative to the other guidelines, the requirement for practice and feedback with the rating task is most critical. This is one of the defining characteristics of FOR training and has been shown to be a necessary component of rater training to ensure training transfer. Ideally, this practice should include rating the videotaped performance of crews flying the event sets in the LOS scenario that will be rated by pilot instructors in the future. If available, these tapes should display actual crews, as opposed to scripted crews, because actual crew performance typically contains more subtle variations that are hard for raters to observe and distinguish. Ideally, pilot instructor rater training should consist of practice videos that display a range of crew performance levels (e.g., excellent performance, good performance, and poor performance) on the CRM skills to be assessed. For the purpose of this guideline, a minimum of at least three practice videotapes displaying excellent, average, and poor crew performance is recommended. However, the specific number and types of practice tapes that should be included to ensure the highest probability of training transfer has yet to be determined empirically.

Guideline 5. Feedback should compare pilot

instructor practice ratings to gold standards.

Because of the strong empirical support for FOR training[7,17], we advocate that pilot instructor rater training include feedback based upon gold standards that are developed by expert pilot instructors. In addition, feedback should include information on expert rationales for each gold standard. Specific methodologies for developing gold standards have been presented in the literature[14] and an example of what a gold standard might look like appears in Table 2.

The research suggests that gold standards are imperative to get pilot instructors to rate aircrew CRM skills like expert instructors.

Table 2. Gold standard example.

LOS EVENT SET 3
TRIGGER: System malfunction during climb-out. The malfunction is the LE Slat fails to retract in icing conditions.

EVENT SET GRADES	GOLD STANDARD RATINGS	GOLD STANDARD RATIONALES
CREW CRM	Standard	♦ Crew CRM behaviors that were observed: - The crew requested time on RWY for engine run-up. - The CAPT watched outside the aircraft for sliding during engine run-up while the F/O set throttles to 70%. - The F/O verbalized a plan for handling the LE Slat problem. - The CAPT suggested that the crew wait to deal with the LE Slat problem until the aircraft was on its assigned heading. - The CAPT handled the LE Slat Transit Light – On checklist while the F/O flew and talked to ATC.

Furthermore, by using the same gold standards across pilot instructor training classes to provide feedback, as opposed to norming instructors to the standards established in each training class (i.e., as done in IRR training), greater reliability and accuracy should be observed across classes. Overall, this approach should improve the quality of LOS training during LOFT and SPOT and the accuracy of CRM evaluations made during LOE.

SUMMARY AND CONCLUSIONS

In closing, the guidelines presented here for training pilot instructors to evaluate CRM skills were primarily based on FOR training, which has been shown to be highly effective

in other similar performance assessment contexts. These guidelines are meant to provide instructional designers with strategies for developing pilot instructor rater training in the future. Although we are confident that these guidelines will be effective at increasing the accuracy and reliability of pilot instructors who evaluate CRM skills in LOS scenarios, several questions remain unanswered. Researchers and practitioners in the airline industry need to work together to answer these questions to gain a better understanding of how best to train pilot instructors. Specifically, questions remain regarding: (a) how much practice and feedback is required on the LOS rating task, (b) to what extent does pilot instructor rater training generalize from one LOS scenario other similar scenarios; and (c) how often should recurrent pilot instructor rater training be required. These are some of the obvious theoretical and practical questions that need study, although other questions of interest will undoubtedly arise as pilot instructor rater training programs are developed, applied, and tested for the purposes of assessing and evaluating CRM skills.

AUTHOR'S NOTE

This research was supported by a grant from the NASA Ames Research Center. The views presented in this paper are those of the author(s) and should not be construed as an official NASA position, policy or decision, unless so designated by other official document.

REFERENCES

1. Helmreich, R. L., & Foushee, H. C., (1993). Why crew resource management? Empirical and theoretical bases of human factors training in aviation. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.). *Cockpit resource management,* New York: Academic Press.
2. Office of the Federal Register, (2 October 1990), *Special Federal Aviation Regulation 58 – Advanced Qualification Program*, Federal Register, Vol. 55, No. 91, Rules and Regulations (pp. 40262-40278) Washington, DC: National Archives and Records Administration.
3. Office of the Federal Register, (2 October 1990), *Advanced Qualification Program – Advisory Circular*, Federal Register, Vol. 55, No. 91, Rules and Regulations (pp. 40279-40352) Washington, DC: National Archives and Records Administration.
4. Hamman, W. R., Seamster, T. L., Smith, K. M., & Lofaro, R. J. (1991). The future of LOFT scenario design and validation. *Proceedings of the 6th International Symposium on Aviation Psychology*, 589-594.
5. Birnbach, R. A., & Longridge, T. M. (1993). The regulatory perspective. In E. L. Wiener, B. G. Kanki, & R. L. Helmreich (Eds.), *Cockpit resource management*. New York: Academic Press.
6. Williams, D., Holt, R., & Boehm-Davis, D. (1997). Training for inter-rater reliability: Baselines and benchmarks. *Proceedings of the 9th International Symposium on Aviation Psychology*, 514-520.
7. Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.
8. Woehr, D. J., & Feldman, J. M. (1993). Processing objective and question order effects on the causal relation between memory and judgment in performance

appraisal: The tip of the iceberg. *Journal of Applied Psychology, 78,* 232-241.
9. DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33,* 360-396.
10. Thornton, G. C. & Zorich, S. (1980). Training to improve observer accuracy. *Journal of Applied Psychology, 65,* 351-354.
11. Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology, 73,* 68-73.
12. Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205-212.
13. Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology, 79,* 525-534.
14. Baker, D. P., Swezey, R. W., & Dismukes, R. K. (1998). *A methodology for developing gold standards for rater training videotapes*. Washington, DC: Federal Aviation Administration, Office of the Chief Scientific and Technical Advisor for Human Factors.
15. Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73,* 497-506.
16. Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences,* rev. ed. New York: Academic Press.
17. Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Journal, 11,* 22-40.
18. George Mason University (1996). *Developing and Evaluating CRM Procedures for a Regional Air Carrier, Phase I Report*. Washington, DC: Federal Aviation Administration, Office of the Chief Scientific and Technical